Journal of Research of the National Bureau of Standards
Volume 91, Number 1, January-February 1986

Ruggedness Testing—Part I: Ignoring Interactions

Robert C. Paule, George Marinenko, Melissa Knoerdel, and William F. Koch

National Bureau of Standards, Gaithersburg, MD 20899

Accepted: August 29, 1985

A straightforward explanation of the statistical technique of ruggedness testing is presented. Efficient Plackett-Burman designs are used in ruggedness tests. These designs involve the simultaneous change of levels of a number of variables. The designs allow the ruggedness test user to determine the effect of the separated variables on the measurement process. This paper (Part I) deals with the common situation where two-factor and higher order interactions can be safely ignored. A method is presented for evaluating the experimental uncertainties. A detailed example of glass electrode measurements of pH of dilute HCl solutions is used to illustrate ruggedness testing procedures.

Key words: interactions; main effects; orthogonal designs; pH measurements; Plackett-Burman designs; ruggedness tests.

Introduction

The purpose of a ruggedness test is to find the factors that strongly influence measurement results, and to determine how closely one needs to control these factors. Ruggedness tests do not determine optimum conditions for a test method.

In the testing of a protocol, it is frequent occurence that the coordinating scientist is dismayed by the large variabilities observed between different laboratory results. The scientist may have developed the protocol being tested and has taken great care and pride in that development. His laboratory has documented "proof" of high precision and accuracy for the method. What has gone wrong? How can the other laboratories get such wild results?

A large part of the answer may be that the coordinating scientist has been unrealistically consistent in his own laboratory work. He may have always used fixed equipment such as a furnace that was set at 60.0 °C and that did not vary by more than ±0.5 °C. Even though the furnace dial read 60.0 °C, the furnace temperature may in reality have been 64.2±0.5 °C. The constant bias of 4.2 °C did not affect his precision, but it may have affected his accuracy. Other constant errors will, likewise, not affect his precision. In regard to accuracy, these additional errors may partially cancel each other. It is the nature of protocol development that work will continue until the errors do cancel, and the "right" answer is obtained. Thus, the laboratory that has developed the protocol will eventually show both good precision and accuracy. In an interlaboratory experiment, however, conditions are different. The other (individual) laboratories do not have the same biases, and the rather complete cancelling of systematic errors does not occur. Differences in laboratory conditions can result in

About the Authors: Robert C. Paule is a physical scientist assigned to the NBS National Measurement Laboratory (NML). George Marinenko and William F. Koch are chemists in NML's Inorganic Analytical Research Division in which Melissa Knoerdel, a student, serves the Division during summer vacations.

large variabilities between different laboratory results. In frustration, the coordinating scientist may tighten the protocol specifications. One can see that if temperature is important, then even a tightened protocol specification of 60.0±0.1 °C will not be effective unless the biases *between* laboratories are eliminated. A true temperature of 60.0±0.5 °C may be quite satisfactory, but large biases cannot be tolerated.

To work towards perfecting a test method one must first determine if a factor such as temperature is important, and then decide if a true ±0.5 °C tolerance is acceptable. Such matters are best investigated in a single laboratory rather than in multiple laboratories since, here, we are interested in the effect of *changes* in temperature. A constant bias within a single laboratory will not interfere in the investigation of changes of temperature. Other factors associated with the protocol must also be evaluated. How do we proceed?

The coordinating scientist may believe that the protocol contains seven factors (variables) that could influence the measurement results. Suppose it is decided to investigate the effect of each factor at only two levels: at a high level and at a low level. A full factorial investigation of the seven factors at each of the two levels would require $2^7 = 128$ measurements, and this does not include replicate measurements. Fortunately, one does not have to make this many measurements. One can use a class of experimental designs called Plackett-Burman designs [1].[1] It is possible, by using these designs, to study up to N−1 factors using only N measurements.

A Mathematical Model

A brief review of a mathematical model used to describe a measurement result may be helpful in understanding details that are associated with the use of Plackett-Burman designs. For simplicity, consider an experiment with only three factors at each of two levels (eight measurements).

$$Y_{ijk} = Y... + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk}$$

where

Y_{ijk} = a single measured value (i,j,k = 1,2 — the low and the high levels)

$Y...$ = the overall average for all eight measurements

A_i, B_j, C_k = the estimated main effects (the main factors affecting the measurement results)

AB_{ij}, AC_{ik}, BC_{jk} = the estimated two-factor interactions (systematic effects not explained by the main effects)

ABC_{ijk} = the estimated three-factor interactions (systematic effects not explained by the main effects and the two-factor interactions).

There are some restrictions on the main effects and interaction terms in the model. The restrictions will not be given here since they only have to do with the "centering of the data" for the evaluation of the terms. In ruggedness testing we do not center the data about some midpoint, but rather redefine the effects as differences between the results at the high and at the low levels. We will also do away with the subscripts of the above model. We simply recognize that measurement results are affected by various main effects and interactions.

From the general mathematical model one can infer that experiments with a larger number of factors will have a very large number of higher-order interactions. It is generally believed that main effects tend to be most important in describing (or controlling) the measurement results, that two-factor interactions are less important, and that higher order interactions are even less important. Plackett-Burman designs are well suited for measurement processes that have negligible interactions.

Use of Plackett-Burman Designs

The most common use of Plackett-Burman (PB) designs with N measurements allows one to get the most important (main effects) information. With N measurements, however, the N−1 main effects are confounded with the two-factor and with higher order interactions. If the interactions are relatively small, then we may be satisfied in making only N measurements and obtaining slightly contaminated estimates for the N−1 main effects. Experience has tended to show that one gains more useful information by examining additional factors than by evaluating the interactions.

Numerous PB-designs are available [1]. A PB-design for seven factors and eight measurements is given in table 1. A (+) for a given factor indicates that the measurement is made with that factor set at the high level, and a (−) indicates the factor is to be at the low level. All seven factors are set for each measurement and a single result is obtained from each of the eight measurements. The measurements should be made in a random order. Typical measurement results are shown at the far right of the design. Scanning down each column of the design one sees that there are equal numbers of (+) and (−) factor settings.

[1] Figures in brackets indicate literature references.

Table 1. A Plackett-Burman design for N=8.

	Factor							
Run	*A*	*B*	*C*	*D*	*E*	*F*	*G*	Results
1	+	+	+	−	+	−	−	1.1
2	−	+	+	+	−	+	−	6.3
3	−	−	+	+	+	−	+	1.2
4	+	−	−	+	+	+	−	0.8
5	−	+	−	−	+	+	+	6.0
6	+	−	+	−	−	+	+	0.9
7	+	+	−	+	−	−	+	1.1
8	−	−	−	−	−	−	−	1.4

The effect of any factor such as A, for example, is simply calculated as the average of the measurements made at the high level minus the average of the measurements made at the low level.

Effect of A

$$=\Sigma\frac{A(+)}{N/2}-\Sigma\frac{A(-)}{N/2}=2/N\times[\Sigma A(+)-\Sigma A(-)]\,. \quad (1)$$

Effect of A

$$=2/8\times[(1.1+0.8+0.9+1.1)-(6.3+1.2+6.0+1.4)].$$
$$=-2.75$$

The PB-design (see table 1) is constructed such that the $\Sigma A(+)$ and the $\Sigma A(-)$ terms will *each* contain an equal number of $B(+)$ and $B(-)$ terms. Thus, the A effect is orthogonal, i.e., is not affected by the B effect. In the PB-designs all main effects (columns) are orthogonal to all other main effects (columns). This orthogonality, however, does not extend to the interactions. The orthogonality of the main effects and the acceptance of a slight contamination of estimates for the main effects (by the interactions) are the major characteristics of ruggedness testing. For many practical problems this is all that is needed.

For the PB-design, the standard deviation for an effect, such as A, is obtained by using eq (1) and the standard deviation of a single measurement σ.

$$\sigma_{\text{effect }A}=\sqrt{(4/N^2)\times\text{Var}\,[\Sigma A(+)-\Sigma A(-)]}$$
$$=\sqrt{(4/N^2)\times N\sigma^2}$$
$$\sigma_{\text{effect }A}=2\sigma/\sqrt{N} \quad (2a)$$

The same equations for the PB-design apply when the standard deviation σ is replaced by a sample estimate, s.

$$s_{\text{effect }A}=2s/\sqrt{N} \quad (2b)$$

Two methods for determining a sample estimate of the standard deviation of a single measurement, s, will be presented.

PB-Design Considerations

Equation 2b shows that the standard deviation of an effect is inversely proportional to $\sqrt{N}$, the number of measurements made. One is therefore tempted to use large PB-designs. Practical experience, however, favors moderate size designs. Overly large designs require the correct setting of too many factors, and this increases the chance for blunders. In addition, large designs require more time to complete and one becomes concerned that other factors not being considered in the design can change and distort the results. The effects of incorrect factor settings and of shifting experimental conditions are propagated into *all* of the calculated results (see eq 1). The above listed ($N=8$) PB-design is a suitable size for most experiments. If more factors need to be studied, they can be handled by using a second ($N=8$) PB-design. This latter procedure may even involve the repeated testing of some of the more important factors from the first design. The ($N=8$) PB-design can also be conveniently used to study two-factor interactions (see Ruggedness Testing—Part II: Recognizing Interactions).

In general, the size of all effects in a PB-design will increase with increased separation of the high and low factor settings. We have implicitly assumed that the main effects are linear. It seems prudent to only use moderate separations of the high and low settings so that the measured effects will be relatively linear and, at the same time, large relative to the measurement error. For the high and low settings of the factors it is suggested that one use the extreme limits that one may expect to observe between different qualified laboratories.

Judging the Effects

How can one judge if any of the estimated main effects are too large? Since the main effects are expressed in the units of the measurement, one can simply make a direct judgment whether the change associated with a factor shift from a high level to a low level is too large, or not. Other, more quantitative methods of judgment which analyze the variance of measurements are given below. We should recognize that these quantitative methods still only give tentative answers and that follow-up or confirmatory experiments are frequently needed.

If n auxiliary replicate measurements are available, one can estimate the within-laboratory measurement variability, s. A t-test (with $n-1$ degrees of freedom) can be used to judge if a main effect is statistically significant relative to the measurement variability. Note that the n from the auxiliary replicate measurements will not generally be the same as the N of the ruggedness test.

$$t_{n-1}=\frac{\text{effect } A}{S_{\text{effect } A}}$$

Using eq 2b, this t-test can be written in the following form:

$$t_{n-1}=\frac{\text{calculated } A}{2s/\sqrt{N}}. \qquad (3)$$

Action should be taken if the effect of a factor is statistically significant, and if the size of the effect is of practical importance; we should then tighten the protocol specification for that factor. This will help reduce the interlaboratory variability.

One may wish to repeat the complete PB-experiment so as to obtain better estimates of the factors and to get a current estimate of the within-laboratory measurement variability, s. In estimating the measurement variability one needs to guard against the occurrence of a possible measurement shift between the running of the two designs. This can be handled mathematically. Let us now work through a real example.

This ruggedness testing example deals with factors that may influence the determination of the pH in dilute acid solutions when measurements are made by use of a glass electrode. Table 2 gives the seven factor (N=8) PB-design which was used. This convenient design was first suggested by F. Yates [2]. It was frequently used by W. J. Youden [3] who did much of the pioneering work in ruggedness testing.

The above Yates-Youden design can be obtained from the seven-factor PB-design of table 1 by relabelling the PB-columns A-G to read C, F, G, D, E, B, A, and the PB-rows 1-8 to read 2, 3, 5, 4, 7, 8, 6, and 1. One then rearranges the columns and rows to be in the usual alphabetic and numeric order. The above operations are perfectly acceptable since the assignment of column and row labels is arbitrary and the rearrangement of the columns and rows has no effect on the overall arithmetic operations. Such rearrangements are, in fact, one means of randomizing the assignment of variables.

Table 2. The seven-factor PB design.

	Factor						
Run	A	B	C	D	E	F	G
1	−	−	−	−	−	−	−
2	−	−	+	−	+	+	+
3	−	+	−	+	−	+	+
4	−	+	+	+	+	−	−
5	+	−	−	+	+	−	+
6	+	−	+	+	−	+	−
7	+	+	−	−	+	+	−
8	+	+	+	−	−	−	+

A number of pH measurement experiments were run using six different dilute acid solutions. For simplicity of presentation, Part I discusses only the results from one of the solutions, an HCl solution with a known pH of 2.985. Subjects of more involved PB-testing and comparisons between the different acid solutions are described in Part II. The seven factors that were studied are listed below. The first listed level for each factor has been arbitrarily assigned the positive sign in the above table.

A. Temperature: 25 °C or 30 °C.
B. Stirring during the pH measurement: Yes or No
C. Dilution (0.5 mL distilled H_2O/20 mL of solution:
Yes or No
D. Depth of electrode immersion: 1 cm or 3 cm below liquid surface
E. Addition of $NaNO_3$ (0.033 mol/L of solution): Yes or No
F. Addition of KCl (0.067 mol/L of solution): Yes or No
G. Electrode equilibration time before reading the pH: 10 or 5 minutes

The above is only a partial list of factors that will change the observed value of the pH. Obviously, all other factors that are not listed above need to be kept constant. The particular, constant levels of these other factors will result in some specific offset in the pH measurements. In the ruggedness test, however, this fixed offset need not concern us since we are only interested in the measurement changes (the effects) that occur when the above seven factors ($A-G$) are changed.

Results from the ruggedness test are given in table 3. The complete experiment was also repeated on a second day. A different random order of measurement was used for each day. The two sets of measurement results are given at the far right of the design.

For the first set of the above reported measurements, the effect of factor A is calculated from eq 1 as the difference of the average value when 25 °C is used and the average value when 30 °C is used, i.e., $(2999+3055+3049+2949)/4-(2904+3015+3006+2964)/4=3013-2972=+41$. The averages and differ-

Table 3. Design and test results.

Factor							Observed pH (milli-pH units)	
A	*B*	*C*	*D*	*E*	*F*	*G*		
30	N	N	3	N	N	5	2904	2895
30	N	Y	3	Y	Y	10	3015	3017
30	Y	N	1	N	Y	10	3006	2990
30	Y	Y	1	Y	N	5	2964	2935
25	N	N	1	Y	N	10	2999	2983
25	N	Y	1	N	Y	5	3055	3053
25	Y	N	3	Y	Y	5	3049	3044
25	Y	Y	3	N	N	10	2949	2949
						Average	2993	2983

ences of the averages (the effects) are given for factors $A-G$ in the third and fourth columns of table 4. Similar calculations for the second set of measurements are given in the fifth and sixth columns of the table.

Testing the Effects From Repeated (pH) Experiments

Generally good agreements are observed between the calculated effects from the two sets of measurements. Effects A, D, E, and F are relatively large and are of interest. The average C effect is $(6+11)/2=+8.5$. To help decide if the C effect value is real, or if it might simply be due to imprecisions in the measurements, let us make a t-test.

$$t=\frac{\text{effect of avg. } C}{S_{\text{effect of avg. } C}}$$

Table 4. The effects for factors $A-G$. (milli-pH units)

		First Data Set		Second Data Set		Differences (d)
Factor	Level	Average	Effect	Average	Effect	betw. effects
A	25	3013		3007		
A	30	2972	+41	2959	+48	−7
B	Y	2992		2980		
B	N	2993	−1	2987	−7	+6
C	Y	2996		2989		
C	N	2990	+6	2978	+11	−5
D	1	3006		2990		
D	3	2979	+27	2976	+14	+13
E	Y	3007		2995		
E	N	2979	+28	2972	+23	+5
F	Y	3031		3026		
F	N	2954	+77	2941	+85	−8
G	10	2992		2985		
G	5	2993	−1	2982	+3	−4

Since the estimate for each effect is now the average of two experiments the t-test, derived in the form of eq 3, must be modified as follows:

$$t=\frac{\text{calculated avg. } C}{2s/\sqrt{2N}}. \tag{4}$$

The estimate of the standard deviation, s, and the associated degrees of freedom for the t-test are obtainable from our measurements. Since the two sets of measurements were run on different days, we should be concerned that one set of measurements could be offset relative to the other set. Let us therefore calculate the s value by a method that is not vulnerable to an offset between the two sets of measurements.

We first note that an offset between the sets of measurements will not influence the values of the calculated effects. Let us therefore consider the differences between the effects as calculated for the above example (see table 4, column 7). Since we are considering the same effects from the two sets of experiments, the statistically expected values of the differences between the effects are zero. The variance of the difference is therefore the expected value of the squared differences.

$$\text{Variance of } (d)=\text{Expected value of } (d^2)$$

$$\approx \Sigma d^2/(N-1) \tag{5}$$

An estimate of the expected value of (d^2) is obtained by simply averaging the squares of the differences listed in table 4, column 7. Our calculated estimate is $384/7=54.9$.

We next note that the variance of the difference (between the duplicated effects) is the sum of the variances of the two effects. The variances of the two effects should be the same since the two sets of experiments were done in the same laboratory. Equation 2b described the sample estimate for the square root of the variance of an effect. Therefore:

$$\text{Estimated variance of } (d)=4s^2/N+4s^2/N=8s^2/N. \tag{6}$$

By combining eqs (5) and (6) and rearranging we obtain an estimate of the standard deviation of a single measurement that has $N-1$ degrees of freedom associated with it.

$$s=\sqrt{[\Sigma d^2/(N-1)]\times N/8} \tag{7}$$

The desired t-test is obtained by combining eqs 4 and 7.

$$t_{N-1}=\frac{\text{calculated avg. } C}{2\sqrt{[\Sigma d^2/(N-1)]\times N/8}/\sqrt{2N}} \quad (8)$$

In the current example, N equals eight so we get:

$$t_7=\frac{\text{calculated avg. } C}{2\sqrt{\Sigma d^2/7}\,/\sqrt{2\times 8}}$$

$$=\frac{+8.5}{\sqrt{384/7}\,/\sqrt{16}}=+2.30\,.$$

This quantity, in absolute value, it is slightly less than the 5% critical t-value of 2.36. It is not quite statistically significant. The C factor describes the effect of a small dilution, as one might get from not properly wiping dry the glass electrode.

As mentioned above, if the effect of any factor is too large one may wish to tighten the specification for that factor. The goal, of course, is to reduce the inter-laboratory variability. More detailed discussions of the pH measurement experiments are presented in Part II.

Other PB-Designs

Numerous Plackett-Burman designs [1] are available. The following is a method for constructing the designs for various numbers of measurements, N=4, 8, 12, 16, and 20. The first row of each design is given opposite the N-value. Each row specifies the N−1 high [+] and low (−) factor settings.

N=4	+ + −
N=8	+ + + − + − −
N=12	+ + − + + + − − − + −
N=16	+ + + + − + − + + − − + − − −
N=20	+ + − − + + + + − + − + − − − − − + + −

For any selected N-value, the corresponding set of (+) and (−) signs is written down as the first row of the design. The second row of the design is obtained by copying the first row after shifting it one place to the right and putting the last sign of row 1 in the first position of row 2. This type of cyclic shifting should be done a total of N-2 times, after which a final row of all minus signs is added. The result of this procedure for the N=8 Plackett-Burman design is given in table 1.

Some ruggedness test studies may not involve exactly N−1 factors. If we believe, for example, that only five instead of seven factors might influence the measured results, we might use two dummy factors. For one of the dummy factors we might pour a solution with our left hand for the (+) level and with our right hand for the (−) level. The calculated "effect" for the dummy factor should be small and should simply reflect our random errors of measurement.

Conclusions

A straightforward explanation of the statistical technique of ruggedness testing has been presented. Orthogonal Plackett-Burman designs allow the ruggedness test user to efficiently evaluate the effects of the separated variables on a measurement process. The present article (Part I) deals with the common situation where two-factor and higher order interactions can be safely ignored.

References

[1] Plackett, R. L., and J. P. Burman, The Design of Optimum Multifactorial Experiments, Biometrika, Vol. 33, 305-325 (1946).

[2] Yates, F., Complex Experiments, J. Roy. Statistical Soc. (Supplement), Vol. 2, 181-247 (1935).

[3] Youden, W. J., Designs for Multifactor Experimentation, Industrial and Engineering Chemistry, Vol. 51, 79A-80A (1959).

[4] Diamond, W. J., Practical Experimental Designs for Engineers and Scientists, pp. 103 and 110, Lifetime Learning Publications, Belmont, CA (1981).

[5] Marinenko, George; Robert C. Paule, William F. Koch, and Melissa Knoerdel, Effect of Variables on pH Measurement in Acid-Rain-Like Solutions as Determined by Ruggedness Tests, J. Res. Natl. Bur. Stand. **91-1** (1986).